MANG1041_Individual_80%.docx

by Nalie Do

Submission date: 10-May-2025 04:53PM (UTC+0100)
Submission ID: 258700672
File name: MANG1041_Individual_80_.docx (26.61K)
Word count: 1489
Character count: 7985

MANG1041

Individual Coursework (80%)

35597836

PART 1: DATA ANALYSIS USING SAS

```
/* MANG1041 Individual Coursework - Las Vegas Hotel Review Analysis */
/* This program analyses hotel review data from Las_Vegas.xlsx to explore insights
   about customer origins, traveler types, hotel ratings, and hotel facilities. */
/* All data steps and procedures are done in the WORK library as required. */

/* TASK 1: Import the data into SAS */
/* Importing Las_Vegas.xlsx and saving it in the WORK library as 'las_vegas'. */
proc import datafile="C:\Users\nad1e23\OneDrive - University of
Southampton\Documents\MANG1041\Las_Vegas.xlsx"
        out=work.las_vegas
        dbms=xlsx replace;
        sheet="Sheet1";
        getnames=yes;
run;

/* TASK 2: Generate frequency of user_country values */
/* This helps identify where most reviewers come from, sorted by most frequent. */
proc freq data=work.las_vegas order=freq;
        tables user_country;
run;

/* TASK 3: Identify top 3 countries by number of reviews */
/* Answer: Based on the output, the top 3 are: USA, UK, Canada. */


/* TASK 4: Compute descriptive statistics for score grouped by traveler_type */
```

```
/* This shows how different traveler types rate hotels. */
proc means data=work.las_vegas n mean stddev min max;
        class traveler_type;
        var score;
run;


/* TASK 5: Identify highest and lowest scoring traveler types */
/* Answer: Based on mean scores: Highest: Friends & Lowest: Business travelers. */


/* TASK 6: Create a new data set 'hotels' with average score per hotel and facility details per
hotel */
/* Some facilities like pool, gym, tennis_court, …. are character variables with values like "Yes"
or "No". Since each hotel has either all "Yes" or all "No" for a given facility (no mixed values),
we can use the max() function to pull the consistent value (either "Yes" or "No") for each hotel.
The max() function works here because it ensures that we get the same "Yes" or "No" across all
reviews for each hotel without any inconsistencies. */
proc sql;
        create table work.hotels as
        select hotel_name,
                mean(score) as avg_score,
                max(pool) as pool,
                max(gym) as gym,
                max(tennis_court) as tennis_court,
                max(spa) as spa,
                max(casino) as casino,
                max(free_internet) as free_internet,
                max(hotel_stars) as hotel_stars,
                max(nr__rooms) as nr__rooms
        from work.las_vegas
```

```
        group by hotel_name;
quit;

/* TASK 7: Identify the best and worst rated hotels based on average score */
/* First, sort hotels by avg_score (descending) to get the best one. */
proc sort data=work.hotels out=work.sorted_hotels;
        by descending avg_score;
run;
/* We only print the top observation (obs=1) for the best-rated hotel because the question asks for one only. */
proc print data=work.sorted_hotels (obs=1);
        title "Best Rated Hotel";
run;
/* Sorting in ascending order to get worst hotel and print the top observation. */
proc sort data=work.hotels OUT=work.sorted_hotels_asc;
   by avg_score;
run;
proc print data=work.sorted_hotels_asc (obs=1);
   title "Worst Rated Hotel";
run;
title; /* Clear the title */
/* Answer: Best Rated Hotel = Wynn Las Vegas, Worst Rated Hotel = Circus Circus Hotel & Casino Las Vegas. */

/* TASK 8: Correlation between hotel_stars and nr__rooms */
/* Checking if bigger hotels (in terms of rooms) are also more luxurious (more stars). */
proc corr data=work.hotels;
        var hotel_stars nr__rooms;
```

```
run;


/* TASK 9: Interpretation of correlation */

/* Answer: The correlation coefficient is r = 0.42097, which shows a MODERATE positive
correlation between hotel stars and number of rooms. This means that hotels with more stars tend
to have more rooms, but the relationship is not very strong. */

/* p-value is 0.0574, which is slightly above the 5% significance level (0.05). This means the
correlation is not statistically significant and we do not have strong evidence that this pattern is
reliable in the larger population. */

/* TASK 10: Build a linear regression model with avg_score as dependent variable */

/* Because 'pool', 'gym', 'spa', and 'free_internet' are character variables ("Yes" or "No"). So, to
use them in regression analysis, they need to be converted to binary numeric (1 = Yes, 0 = No).
*/

data work.hotels_reg;

   set work.hotels;

   pool_num = (upcase(pool) = "YES");  /* Using upcase() ensures the code works even if the
values are later entered as "Yes", "yes", or "YES" (it's safer and more flexible) */

   gym_num = (upcase(gym) = "YES");

   spa_num = (upcase(spa) = "YES");

   free_internet_num = (upcase(free_internet) = "YES");

run;

proc reg data=work.hotels_reg;

        model avg_score = hotel_stars pool_num gym_num spa_num free_internet_num;

run;


/* TASK 11: Evaluate model fit using R-squared */

/* Answer: R-squared = 0.7493: This means that the model explains approximately 75% of the
variability in average hotel ratings. The model fits the data pretty well and that the predictors
```

(hotel_stars, pool_num, gym_num, spa_num, free_internet_num) collectively have a strong explanatory power for avg_score. */

/*Adjusted R-Square = 0.6658: This is slightly lower but still strong. */

/* TASK 12: Create PDF report of hotels with avg_score ≥ 4 */

/* This line starts the PDF file and chooses a professional style for the report. */

ods pdf file="C:\Users\nad1e23\OneDrive - University of Southampton\Documents\MANG1041\hotels.pdf" style=sapphire;

title "The Best Hotels in Las Vegas"; /* This sets the title that will appear at the top of the report. */

/* This prints a table showing only hotel name, stars, and average score. */

proc print data=work.hotels noobs;

    var hotel_name hotel_stars avg_score;

    where avg_score >= **4**;

run;

/* This ends the PDF and saves the file. */

ods pdf close;

/* TASK 13: Create a subset of reviews from one specific group of people only */

/* I chose ‘Friends’ group, so I created a new dataset named ‘selected’ to show the data from Friends group only. */

data work.selected;

    set work.las_vegas;

    where traveler_type = "Friends";

run;

/* Use SQL to calculate the average score for each hotel, based only on the 'Friends' group. */

/* This creates a new table (hotels2) showing hotel_name and its average score from Friends. */

proc sql;

```
        create table work.hotels2 as
        select hotel_name,
                        mean(score) as avg_score
        from work.selected
        group by hotel_name;
quit;

/* TASK 14: Identify best and worst hotels rated by Friends group */
/* Sort descending to find the top-rated ones. */
proc sort data=work.hotels2 out=work.fen_sorted;
        by descending avg_score;
run;
/* Create a table with hotels that have the highest avg_score */
/* Proc sql is used here instead of proc sort (which used in step 7) because proc sql allows SAS
to collect all hotels that match the maximum/minimum score */
proc sql;
   create table work.best_hotels as
   select *
   from work.fen_sorted
   where avg_score = (select max(avg_score) from work.fen_sorted);
quit;
/* Print */
proc print data=work.best_hotels;
   title "Best Rated Hotels by Friends";
run;
/* Next, keep using proc sql and sort ascending to find worst rated hotels. */
proc sort data=work.hotels2 out=work.fen_sorted_asc;
```

```
        by avg_score;
run;
proc sql;
   create table work.worst_hotels as
   select *
   from work.fen_sorted_asc
   where avg_score = (select min(avg_score) from work.fen_sorted_asc);
quit;
/* Print */
proc print data=work.worst_hotels;
   title "Worst Rated Hotels by Friends";
run;
/* Answer:
Five Best Rated Hotels by Friends (average score = 5): Hilton Grand Vacations on the
Boulevard, The Venetian Las Vegas Hotel, Trump International Hotel Las Vegas, Tuscany Las
Vegas Suites & Casino, Wynn Las Vegas. Only one Worst Rated Hotel by Friends (average score
= 3.42857): Circus Circus Hotel & Casino Las Vegas */
```

PART 2: SHORT REPORT

If I had to recommend a tool for business analytics to my future employer, SAS will not be my first choice, but it is not something I will completely ignore. It mainly depends on the type of company and what the employer really requires for their data.

Firstly, there are lots of advantages to using SAS. SAS is solid, reliable, and trustworthy, and it excels at handling large data sets. Its built-in features make it easy to run advanced analysis without a lot of code, which is why it is widely used in industries requiring a high level of security, accuracy, and regulatory compliance like banking (HSBC), healthcare (NHS), education (Pearson), and government agencies. Additionally, one thing I particularly appreciate when using SAS is that it provides a full platform from cleaning data to building models and generating visual results. For companies already using SAS, or those in regulated sectors, I highly recommend sticking with this software, as it delivers exceptional results in the right contexts.

However, the main issue with SAS is that it is expensive. Compared to free open-source tools like Python or R, which are common, easier to learn, and have larger user communities that keep them up to date and offer quick solutions, SAS can feel less friendly and flexible. Many companies today are moving away from SAS because it can be harder to hire people proficient in it, and it also requires complicated training efforts.

In conclusion, I would assess that if a company already uses SAS or operates in a highly regulated environment that requires security and compliance, SAS remains a solid choice, trusted and effective in those settings. But if the answer is no, and the company's goal is to build a modern, cost-effective, and flexible analytics setup, especially for younger, fast-growing companies, I would prioritise recommending Python or R instead.

MANG1041_Individual_80%.docx

GRADEMARK REPORT

FINAL GRADE

94/100

GENERAL COMMENTS

Part 1: 80/80

When using the WORK library, there is no need to use its
name in the code.

Tasks:

T1: 3/3

T2: 2/2

T3: 3/3

T4: 3/3

T5: 2/2

T6: 10/10

T7: 2/2

T8: 3/3

T9: 2/2

T10: 6/6

T11: 4/4

T12: 5/5

T13: 12/12

T14: 3/3

Addresses tasks/ question set (Performing the tasks using a
SAS program): 60/60

Structure (Comments, the structure and elegance of the SAS
program): 20/20

Part 2: 14/20

The report is generally OK. It covers some advantages and
disadvantages of SAS. It could include stronger/more convincing arguments and be more in-depth. There is a clear and thoughtful
recommendation. The presentation quality is high. The report structure could be
improved, e.g. by dividing the contents into sections and subsections.

Analysis (Rigorously argued and focussed analysis): 6/10

Presentation (Clearly presented report): 8/10

Very well done!

Comment 1

Good comment!